Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

КОБИЛІН ІЛЛЯ ОЛЕГОВИЧ

УДК 004.032.26

НЕЧІТКА КЛАСТЕРИЗАЦІЯ ЧАСОВИХ РЯДІВ В ІНТЕЛЕКТУАЛЬНОМУ АНАЛІЗІ ПОТОКІВ ДАНИХ

05.13.23 – системи та засоби штучного інтелекту

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2019

Дисертацією є рукопис.

Робота виконана в Харківському національному університеті радіоелектроніки Міністерства освіти і науки України.

Науковий керівник – доктор технічних наук, професор
Бодянський Євгеній Володимирович,
Харківський національний університет радіоелектроніки, професор кафедри штучного інтелекту.

Офіційні опоненти: доктор технічних наук, професор
Субботін Сергій Олександрович
Запорізький національний технічний університет МОН України, завідувач кафедри програмних засобів;

кандидат технічних наук, доцент
Гороховатський Олексій Володимирович,
Харківський національний економічний університет ім. Семена Кузнеця МОН України, доцент кафедри інформатики та комп'ютерної техніки.

Захист відбудеться «26» вересня 2019 р. о 15-00 годині на засіданні спеціалізованої вченої ради Д 64.052.01 Харківського національного університету радіоелектроніки за адресою: 61166, м. Харків, пр. Науки, 14.

З дисертацією можна ознайомитись у бібліотеці Харківського національного університету радіоелектроніки за адресою: 61166, м. Харків, пр. Науки, 14.

Автореферат розісланий « 29 » липня 2019 р.

Учений секретар
спеціалізованої вченої ради Є.І. Литвинова

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми На сьогоднішній час тенденція обробки великих обсягів інформації та їх аналіз за допомогою методів інтелектуального аналізу даних дає змогу зрозуміти різноманіття процесів для її подальшого використання у сферах життєдіяльності, які супроводжують людину.

Значну частину інформації, пов'язану з обробкою великих обсягів даних, містять часові ряди. Однак, однією з типових проблем обробки часових рядів є як їх нерівномірне квантування (з причини нерівномірного вимірювання реальних сигналів), так і їх багатовимірність.

Дослідження часових рядів та методи їх обробки описані у працях Бездека Дж., Келлер А., Клавонна Ф., Кохонена Т., Хьоппнера Ф., Машталіра В.П, Путятіна Є.П. та інших вчених, але в цих роботах не розглядались питання аналізу несинхронізованих даних та їх кластеризації, при якій виникає ефект концентрації норм, «прокльон розмірності», рядів, які не мають стохастичної природи або зашумлені викидами, нестаціонарності характеристик цих рядів.

Актуальною та не вирішеною у працях вищезазначених авторів залишається задача, в якій часовий ряд обробляється як вибірка загалом, а не окремими спостереженнями, тобто самі спостереження об'єднані у формі пакету і саме у такому вигляді подаються на обробку. Також до недоліків відомих методів можна віднести неможливість обробки невеликих проміжків спостережень та невміння розпізнавати можливі похибки.

Враховуючи сучасні потреби для розв'язання задач ефективного аналізу та нечіткої обробки часових рядів з нерівномірно розподіленими спостереженнями, які не дозволяють використовувати стандартні методи для обробки в онлайн режимі, є доцільною розробка нових методів нечіткої кластеризації таких рядів у рамках концепції інтелектуального аналізу даних.

Таким чином, нечітка кластеризація часових рядів в інтелектуальному аналізі потоків даних є актуальним науковим завданням.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана в рамках держбюджетних НДР: «Нейро-фаззі системи для поточної кластеризації та класифікації послідовностей даних в умовах їх спотворення відсутніми і аномальними спостереженнями» (№ДР 0113U000361); «Динамічний інтелектуальний аналіз послідовностей нечіткої інформації в умовах суттєвої невизначеності на основі гібридних систем обчислювального інтелекту» (№ДР 0116U002539), які виконувалися у Харківському національному університеті радіоелектроніки, згідно наказів Міністерства освіти і науки України за результатами конкурсного відбору проектів наукових досліджень, у яких автор брав участь як виконавець.

Мета та задачі дослідження. Метою роботи є розробка онлайн методів нечіткої кластеризації нерівномірно квантованих асинхронних нестаціонарних часових рядів для підвищення ефективності та скорочення часу обробки в інтелектуальному аналізі потоків даних.

Для досягнення поставленої мети необхідно розв'язати такі наукові задачі:

– проведення аналізу існуючих методів та підходів до кластеризації часових рядів;

– розробка методу передобробки часових рядів, що спотворені аномальними викидами та збуреннями;

– розробка онлайн модифікації методу кластеризації коротких часових рядів;

– розробка методу онлайн кластеризації багатовимірних часових рядів;

– розробка методу онлайн кластеризації асинхронних часових рядів, несхильного до впливу ефекту концентрації норм;

– проведення експериментів на основі тестових та реальних даних.

Об'єкт дослідження – процес інтелектуального аналізу потоку даних у формі часових рядів.

Предмет дослідження – методи інтелектуального аналізу для нечіткої онлайн кластеризації багатовимірних часових рядів з асинхронними тактами квантування, що призначені для аналізу потоків даних.

Методи дослідження: методи адаптивної фільтрації для обробки часових рядів, забруднених аномальними викидами та збуреннями; методи інтелектуального аналізу даних для вирішення задач нечіткої кластеризації рядів; методи нечіткої логіки для вирішення задач фаззі-кластеризації; методи машинного навчання для синтезу онлайн методів кластеризації.

Наукова новизна отриманих результатів:

1. Вперше запропоновано метод кластеризації, який несхильний до ефекту концентрації норм, що дозволяє вирішувати задачу кластеризації в онлайн режимі за умов перетину класів та асинхронних нерівномірно квантованих часових рядів за рахунок використання спеціальної цільової функції нечіткої кластеризації.

2. Вперше запропоновано послідовний онлайн метод кластеризації багатовимірних часових рядів, що базується на апараті гібридних систем обчислювального інтелекту, який дозволив вирішувати задачу кластеризації даних, які послідовно надходять на обробку з нерівномірними тактами квантування.

3. Отримав подальший розвиток метод адаптивної кластеризації, що базується на методах ймовірнісної та можливісної кластеризації коротких часових рядів, які, у свою чергу, засновані на метриці спеціального вигляду, що дозволяє значно спростити чисельну реалізацію методу, за рахунок використання метрики на основі тангенсів кутів нахилу, що на відміну від відомих методів вирішує задачу кластеризації нерівномірно квантованих часових рядів.

4. Отримав подальший розвиток метод робастної адаптивної ідентифікації нестаціонарних часових рядів в онлайн режимі надходження потоку даних, який характеризується простотою обчислювальної реалізації та вирішує задачу обробки даних, що збурені аномальними викидами, за рахунок використання введеної модифікації критерія Гемана–МакКлюра.

Практичне значення отриманих результатів.

Використання запропонованих моделей та методів дозволяє підвищити ефективність застосування сучасних моніторингових систем для вирішення задач кластеризації даних, які послідовно надходять на обробку з нерівномірними тактами квантування, що базуються на апараті гібридних систем обчислювального інтелекту. Реалізований модуль із запропонованими методами підтвердив свою ефективність у

задачах моніторингу медичних даних в онлайн режимі. У такому разі моніторинг медичних даних дозволяє ефективно виявляти аномалії у хворих у режимі реального часу. Результати досліджень впроваджені у ТОВ «Інфобуд», м. Харків (акт впровадження від 03.10.2018) та у ТОВ «Сайтосс», м. Харків (акт впровадження від 06.10.2018). Результати досліджень впровадженні у Харківському національному університеті радіоелектроніки на кафедрі штучного інтелекту в освітній процес з курсу «Нейромережеві методи обчислювального інтелекту».

Особистий внесок здобувача.

Усі положення, що виносяться на захист, основні результати теоретичних та експериментальних досліджень отримані здобувачем особисто. У публікаціях, написаних у співавторстві, автору належать: [1] - запропонований метод нечіткої кластеризації часових рядів з асинхронними тактами квантування з використанням поліноміального фаззіфікатора; [3] - запропонована адаптивна ідентифікація нестаціонарних часових рядів на основі модифікованого робастного критерія Гемана – МакКлюра та ансамбль адаптивних моделей на його основі; [4] - запропонована самоорганізовна нейро-фаззі мережа для нечіткої кластеризації багатовимірних часових рядів; [5] - запропонований метод адаптивної нейро-фаззі ймовірнісної кластеризації багатовимірних потоків даних; [2] - запропонований метод адаптивної нечіткої кластеризації коротких часових рядів; [7] - запропонована цільова функція нечіткої кластеризації на основі сферичної норми; [6] - запропонована модифікована функція Гемана-МакКлюра; [8] - запропонована оцінка відстані між рядами на основі тангенсів кутів нахилу; [9] - запропоновано використання WTM – правила самонавчання для нечіткої кластеризації часових рядів; [10] - запропоновано зважування координат у просторі ознак.

Апробація результатів дисертації. Основні результати дисертаційної роботи доповідалися й обговорювалися на конференціях:

– First International Conference on Data Stream Mining & Processing (23-27 August 2016, Lviv);

– XIV Міжнародній науково-технічній конференції «Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering» (20-24 лютого 2018 р., м. Славське);

– Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту-ISDMCI-2016 (24-28 травня 2016 р., Залізний порт, м. Херсон);

– VIII Міжнародній школі-семінарі «Теорія прийняття рішень» (26 вересня-1 жовтня 2016 р., м. Ужгород);

– Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту - ISDMCI-2017 (22-26 травня 2017 р., Залізний порт, м. Херсон);

– XIX Міжнародному молодіжному форумі «Радіоелектроніка та молодь у XXI столітті» (20-22 квітня 2015 р., м. Харків);

– XX Ювілейному міжнародному молодіжному форумі «Радіоелектроніка та молодь у XXI столітті» (19-21 квітня 2016 р., м. Харків).

Публікації.

За тематикою дослідження опубліковано 12 наукових праць, з них 1 розділ у колективній монографії, що входить до наукометричної бази Scopus; 1 стаття за кордоном, що входить до наукометричної бази SCOPUS; 3 статті у виданнях, які

зазначені в переліках фахових видань України з технічних наук, 7 публікацій у матеріалах конференцій (2 включено до наукометричної бази даних SCOPUS).

Структура та обсяг дисертації. Дисертація складається із вступу, п’яти розділів, висновків, що містять основні результати, списку використаних джерел і додатку. Загальний обсяг дисертації складає 147 сторінок (з них 130 - основного тексту), містить 51 рисунок, 16 таблиць, список використаних джерел, що включає 115 найменувань та займає 11 сторінок, 2 додатки на 6 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** подано загальну характеристику роботи, обґрунтовано актуальність теми дисертаційної роботи, сформульовано мету і задачі дослідження, визначено наукову новизну, практичне значення і впровадження одержаних результатів, розкрито застосовані методи дослідження, зв'язок роботи з НДР, відомості про публікацію і апробацію результатів досліджень, особистий внесок здобувача.

У першому розділі – проаналізовано стан проблеми кластеризації часових рядів та існуючі підходи до її вирішення. Розглянуто різноманітні метрики, які застосовуються у методах кластеризації часових рядів.

На основі аналізу існуючих методів та метрик зроблений висновок про істотні недоліки існуючих методів кластеризації, визначені мета та задачі дослідження роботи, важливість пошуку і вдосконалення відомих підходів до кластеризації часових рядів.

Другий розділ присвячено розвитку методу робастного оцінювання, який необхідний для фільтрації часових рядів, які зашумлені аномальними викидами та збуреннями, що заснований на мінімізації критеріїв, відмінних від квадратичного.

Запропоновано використати модифіковану функцію Гемана – МакКлюра, що породжена функцією щільності розподілу Коші. Модифікована функція дозволяє вирішити задачі нелінійної оптимізації за умов впливу негативних факторів у вигляді «важких хвостів».

Введено градієнтний метод оптимізації у вигляді:

$$\hat{w}(k) = \hat{w}(k-1) + \eta(k)e(k)J_G(k), \tag{1}$$

де $\hat{w}(k)$ – налаштовні ваги;

$\eta(k)$ – параметр кроку навчання;

$e(k)$ – помилка ідентифікації;

$J_G(k)$ – градієнт модифікованої функції Гемана–МакКлюра.

На основі градієнтного методу оптимізації був введений адаптивний метод фільтрації нестаціонарних часових рядів, що забруднені аномальними викидами:

$$\begin{cases} \hat{w}(k) = \hat{w}(k-1) + \dfrac{\left(y(k) - \hat{w}^T(k-1)x(k)\right)J_G(k)}{r(k)}, \\ r(k) = \alpha r(k-1) + \left\|J_G(k)\right\|^2, \ 0 \le \alpha \le 1, \end{cases} \tag{2}$$

де $y(k)$ – контрольований ряд;

$r(k)$ – параметр, що визначає крок навчання;

α – параметр забування.

На рисунку 1 наведено структурну схему налаштовної моделі, навченої за допомогою модифікованої функції Гемана–МакКлюра.

Така схема простіша за класичні алгоритми ідентифікації аномальних збурень та є розширенням на робастний випадок багатьох алгоритмів навчання.

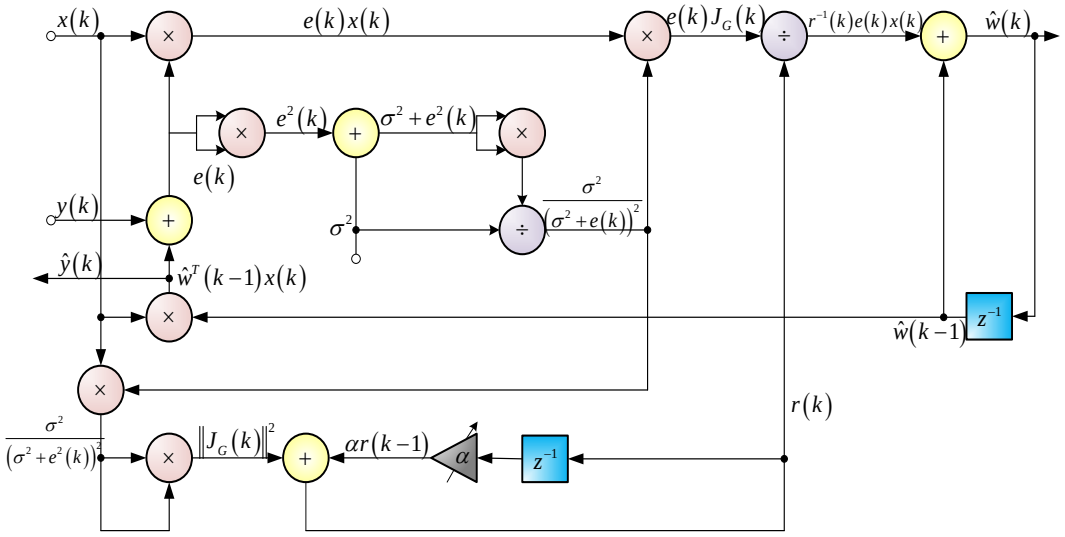

Рисунок 1 – Налаштовна модель, що навчається на основі модифікованої функції Гемана–МакКлюра

У третьому розділі отримав подальший розвиток метод адаптивної кластеризації, що базується на методах ймовірнісної та можливісної кластеризації коротких часових рядів, які, у свою чергу засновані на метриці спеціального вигляду, основою якої є аналіз тангенсів кутів нахилу цих рядів.

Розглянуто кластеризацію коротких одновимірних часових рядів, вихідна інформація про які задана у формі набору вибірок, які містить $N\left(N>n\right)$ часових послідовностей з нерівномірним тактом квантування, що підлягають кластеризації.

При цьому кожна така реалізація може бути представлена у формі $(n\times 1)$ вектора вхідних сигналів у момент дискретного часу $x\left(k\right)=\left(x_1\left(k\right),x_2\left(k\right),...,x_n\left(k\right)\right)^T$.

Нерівномірність квантування означає що $\Delta t_i=t_i-t_{i-1}\neq\Delta t_{i+1}=t_{i+1}-t_i$, тобто $\Delta t_i\neq const$. Тобто, для оцінки відстані між такими вибірками не можуть бути використані ані традиційна евклідова метрика, ані класичні критерії оцінювання.

У зв'язку з цим була розглянута можливість побудови методу кластеризації за особливості, що при взятті першої різниці буде втрачене одне спостереження, тобто усього $\left(n-1\right)$ спостережень $\Delta x_2\left(k\right),\Delta x_3\left(k\right),...,\Delta x_n\left(k\right)$, або, що те ж саме, $tg\alpha_2\left(k\right),tg\alpha_3\left(k\right),...,tg\alpha_n\left(k\right)$, а для його відновлення було введено пакетний (оффлайн) метод нечіткої кластеризації, який є модифікацією алгоритму нечітких *c*-середніх (FCM) на випадок обробки часових рядів з нерівновіддаленими спостереженнями.

Компоненти виразу – це перші різниці дискретного сигналу $x_i(k)$ або тангенси кутів нахилу лінійних функцій, тобто

$$\Delta x_{i+1}(k) = \frac{x_{i+1}(k) - x_i(k)}{\Delta t_{i+1}} = tg\alpha_{i+1}(k). \tag{3}$$

Оскільки внаслідок взяття різниць з ряду видаляється його середнє значення, для відновлення вихідної вибірки за її різницями необхідно доповнити набір цих різниць будь-яким із спостережень вихідної послідовності, наприклад, $x_n(k)$.

Тоді, маючи послідовність різниць $\Delta x_i(k)$, вихідний ряд відновлюється за допомогою співвідношень:

$$\begin{cases} x_{n-1}(k) = x_n(k) - \Delta x_n(k)\Delta t_n, \\ x_{n-2}(k) = x_n(k) - \Delta x_{n-1}(k)\Delta t_{n-1}, \\ \vdots \\ x_1(k) = x_2(k) - \Delta x_2(k)\Delta t_2. \end{cases} \tag{4}$$

Далі, використовуючи методику стандартного нечіткого ймовірнісного кластерного аналізу, шляхом знаходження сідлової точки функції Лагранжа

$$L\left(u_j(k), \tilde{c}_j, \lambda(k)\right) = \sum_{k=1}^{N}\sum_{j=1}^{m} u_j^{\beta}(k)\lambda(k)\left\|\tilde{x}(k) - \tilde{c}_j\right\|^2 + \sum_{k=1}^{N}\lambda(k)\left(\sum_{j=1}^{m} u_j(k) - 1\right), \tag{5}$$

де $u_j(k)$ – рівень належності вектора $\tilde{x}(k)$ до j-го кластера i;

$\tilde{x}(k)$ – вектор спостережень;

$\tilde{c}_j$ – прототип-центроїд кластера;

$u_j^{\beta}(k)$ – функція сусідства, що має форму кошіану;

$\lambda(k)$ – невизначений множник Лагранжа;

$\left\|\tilde{x}(k) - \tilde{c}_j\right\|^2$ – відстань у евклідовій метриці;

m – кількість кластерів, яка встановлюється апріорно;

$\beta > 1$ – параметр фаззіфікації (fuzzifier), який визначає «розмитість» границь між кластерами,

та скориставшись для пошуку сідлової точки (5) рекурентним алгоритмом нелінійного програмування Ерроу-Гурвіца-Удзави, отримуємо адаптивний градієнтний метод нечіткої кластеризації:

$$\begin{cases} u_j(k+1) = \dfrac{\left(\left\|\tilde{x}(k+1) - \tilde{c}_j(k)\right\|^2\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(\left\|\tilde{x}(k) - \tilde{c}_j\right\|^2\right)^{\frac{1}{1-\beta}}}, \\ \tilde{c}_j(k+1) = \tilde{c}_j(k) + \eta(k)u_j^{\beta}(k+1)\left(\tilde{x}(k+1) - \tilde{c}_j(k)\right), \end{cases} \tag{6}$$

де $u_j^{\beta}(k+1)$ відповідає функції сусідства, яка має форму кошіана замість традиційного гауссіана.

При $\beta = 0$ приходимо до стандартного принципу «Переможець отримує все» (WTA), мінімізуючого цільову функцію:

$$E(\tilde{c}_j) = \sum_k \left\| \tilde{x}(k) - c_j \right\|^2,$$

при цьому при $\eta(k) = (k+1)^{-1}$ отримуємо процедуру стохастичної апроксимації:

$$\tilde{c}_j(k+1) = \tilde{c}_j(k) + \frac{1}{k+1}\left(\tilde{x}(k+1) - \tilde{c}_j(k)\right),$$

що веде до стандартної оцінки середнього арифметичного як центроїда.

Методи, пов'язані з оптимізацією лагранжіана (5), мають істотний недолік, пов'язаний з необхідністю виконання обмеження:

$$\sum_{j=1}^{m} u_j(k) = 1. \tag{7}$$

Тому альтернативою ймовірнісним алгоритмам кластеризації є можливісні методи, що пов'язані з мінімізацією цільової функції:

$$E(u_j, \tilde{c}_j, \mu_j) = \sum_{j=1}^{m} \mu_j \sum_{k=1}^{N} \left(1 - u_j(k)\right)^{\beta} + \sum_{k=1}^{N}\sum_{j=1}^{m} u_j^{\beta}(k) \left\| \tilde{x}(k) - \tilde{c}_j \right\|^2, \tag{8}$$

де $\mu_j > 0$ визначає відстань від $\tilde{x}(k)$ до $\tilde{c}_j$, на якій рівень належності приймає значення $0{,}5$, тобто $u_j(k) = 0{,}5$ при

$$\left\| \tilde{x}(k) - \tilde{c}_j \right\|^2 = \mu_j. \tag{9}$$

Адаптивний варіант можливісного методу (11) може бути отриманий в наслідок градієнтної оптимізації цільової функції:

$$d_{STS}^2\left(x(k), x(l)\right) = \left\| \tilde{x}(k) - \tilde{x}(l) \right\|^2, \tag{10}$$

а її оптимізація по $u_j(k)$, $\tilde{c}_j$ та μ_j веде до рекурентного онлайн методу у вигляді:

$$\begin{cases} u_j(k) = \left(1 + \left(\dfrac{\left\| \tilde{x}(k) - \tilde{c}_j(k) \right\|^2}{\mu_j(k)} \right)^{\frac{1}{\beta - 1}} \right)^{-1}, \\ \mu_j(k+1) = \dfrac{\sum\limits_{k=1}^{N} u_j^{\beta}(p) \left\| \tilde{x}(p) - \tilde{c}_j(k) \right\|^2}{\sum\limits_{p=1}^{k+1} u_j^{\beta}(p)}, \\ \tilde{c}_j(k+1) = \tilde{c}_j(k) + \eta(k) u_j^{\beta}(k+1)\left(x(k+1) - \tilde{c}_j(k)\right). \end{cases} \tag{11}$$

Перевагами запропонованої модифікації є можливість кластеризації одновимірних часових рядів за допомогою комбінації методів ймовірнісної та можливісної нечіткої кластеризації, що дозволяє обробляти часові ряди в онлайн режимі.

Четвертий розділ присвячений розробці послідовного онлайн методу кластеризації багатовимірних часових рядів з нерівномірними тактами квантування та методу кластеризації асинхронних нерівномірно квантованих часових рядів за рахунок використання спеціальної цільової функції.

Об'єктом кластеризації, у задачах обробки багатовимірних часових рядів, є припущення, що вихідна інформація подається у вигляді $(q\times n)$-вимірних матриць $X(k)=\{x_{ip}(k)\}$ (тут $i=1,2,...n$, n – номер окремого спостереження q-вимірної послідовності в k-й реалізації (вибірці), $k=1,2...N$, $p=1,2,...,q$ – p-а координата багатовимірного процесу), що містить N $(N>n)$ q-вимірних реалізацій з нерівномірним тактом квантування.

При цьому p-а компонента $X(k)$ може бути представлена у вигляді $(1\times n)$-вектору $x_p(\mathrm{k})=(x_{1p}(k),x_{2p}(k),\ldots,x_{np}(k))$. Тому увесь часовий ряд, що подається на обробку, може бути записаний у вигляді $(q\times n)$ – матриці:

$$\tilde{X}(k)=\begin{pmatrix} \Delta x_{11}(k) & \Delta x_{21}(k) & \cdots & \Delta x_{nl}(k) & x_{nl}(k) \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \Delta x_{ip}(k) & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \Delta x_{1q}(k) & \Delta x_{2q}(k) & \cdots & \Delta x_{nq}(k) & x_{nq}(k) \end{pmatrix}. \qquad (12)$$

При використанні методики нечіткого ймовірнісного кластерного аналізу введемо до розгляду цільову функцію:

$$E(u_j(k),\tilde{C}_j)=\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^{\beta}(k)D_{PS}^2(\tilde{X}(k),\tilde{C}_j)= \\ =\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^{\beta}(k)Tr(\tilde{X}(k)-\tilde{C}_j)(\tilde{X}(k)-\tilde{C}_j)^T \qquad (13)$$

за наявності стандартних обмежень: $\sum_{j=1}^{m}u_j(k)=1$ або $\sum_{j=1}^{m}u_j(k)-1=0$, $k=1,2,...N, 0<\sum_{j=1}^{m}u_j(k)<N$, $j=1,2,...,m$, де $u_j(k)$ – рівень належності матриці $\tilde{X}(k)$ j-му кластеру з матричним центроїдом $\tilde{C}_j$, m – кількість кластерів, що задається апріорно.

Результатом кластеризації є $(N\times m)$ – матриця $U=\{u_{ij}(k)\}$, що має назву матриці нечіткого розбиття, та m матриць-центроїдів $\tilde{C}_j$, $j=1,2,...,m$.

Записавши функцію Лагранжа на матричний випадок та розв'язавши систему рівнянь Каруша-Куна-Таккера, приходимо до методу кластеризації багатовимірних часових рядів у вигляді:

$$\begin{cases} u_j(k) = \dfrac{(Tr(\tilde{X}(k)-\tilde{C}_j)(\tilde{X}(k)-\tilde{C}_j)^T)^{\frac{1}{1-\beta}}}{\sum\limits_{g=1}^{m}(Tr(\tilde{X}(k)-\tilde{C}_g)(\tilde{X}(k)-\tilde{C}_g)^T)^{\frac{1}{1-\beta}}}, \\ \lambda(k) = -\left(\sum\limits_{g=1}^{m} \beta Tr(\tilde{X}(k)-\tilde{C}_g)(\tilde{X}(k)-\tilde{C}_g)^{\frac{1}{1-\beta}}\right)^{1-\beta}, \\ \tilde{C}_j = \dfrac{\sum\limits_{k=1}^{N} u_j^{\beta}(k)\tilde{X}(k)}{\sum\limits_{k=1}^{N} u_j^{\beta}(k)}. \end{cases} \quad (14)$$

Запропонований метод дає змогу вирішити задачі нечіткої кластеризації багатовимірних часових рядів з нерівномірним тактом квантування, що надходять на обробку в онлайн режимі.

Запропоновано метод, при роботі якого оцінюються параметри фаззіфікації та зважування, що виділяють тільки найбільш інформативні значення, тобто таким чином розв'язана задача кластеризації багатовимірних нерівномірно квантованих часових рядів з асинхронним тактом квантування.

Розроблено метод відновлення центроїдів вихідних даних $\tilde{C}_j$ у випадку необхідності визначення нерівномірності квантування.

При цьому кожна вибірка представлена в формі $\left(n(k)\times 1\right)$ – вектора

$$x(k) = \left(x_1(k), x_2(k), ..., x_{n(k)}(k)\right)^T,$$

де кожне спостереження $x_{i(k)}(k)$ реалізовано в момент часу $0 \le t_{i(k)}(k) \le T$.

На рисунку 2 показані реалізації $x(1) = \left(x_1(1), x_2(1), ..., x_{n(1)}(1)\right)^T$, $x(k) = x_1(k), x_2(k), ..., x_{n(k)}(k))^T$ та $x(N) = x_1(N), x_2(N), ..., x_{n(N)}(N))^T$, при цьому у загальному випадку $t_1(1) \ne t_1(k) \ne t_1(N)$ и $t_{n(1)}(1) \ne t_{n(k)}(k) \ne t_{n(N)}(N)$.

Інтервал спостереження всього набору даних може бути заданий у вигляді:

$$\left[t_1 = t_{1\min} = \min\left\{t_1(k)\right\} - t_n = T = \max\left\{t_{n(k)}(k)\right\}\right].$$

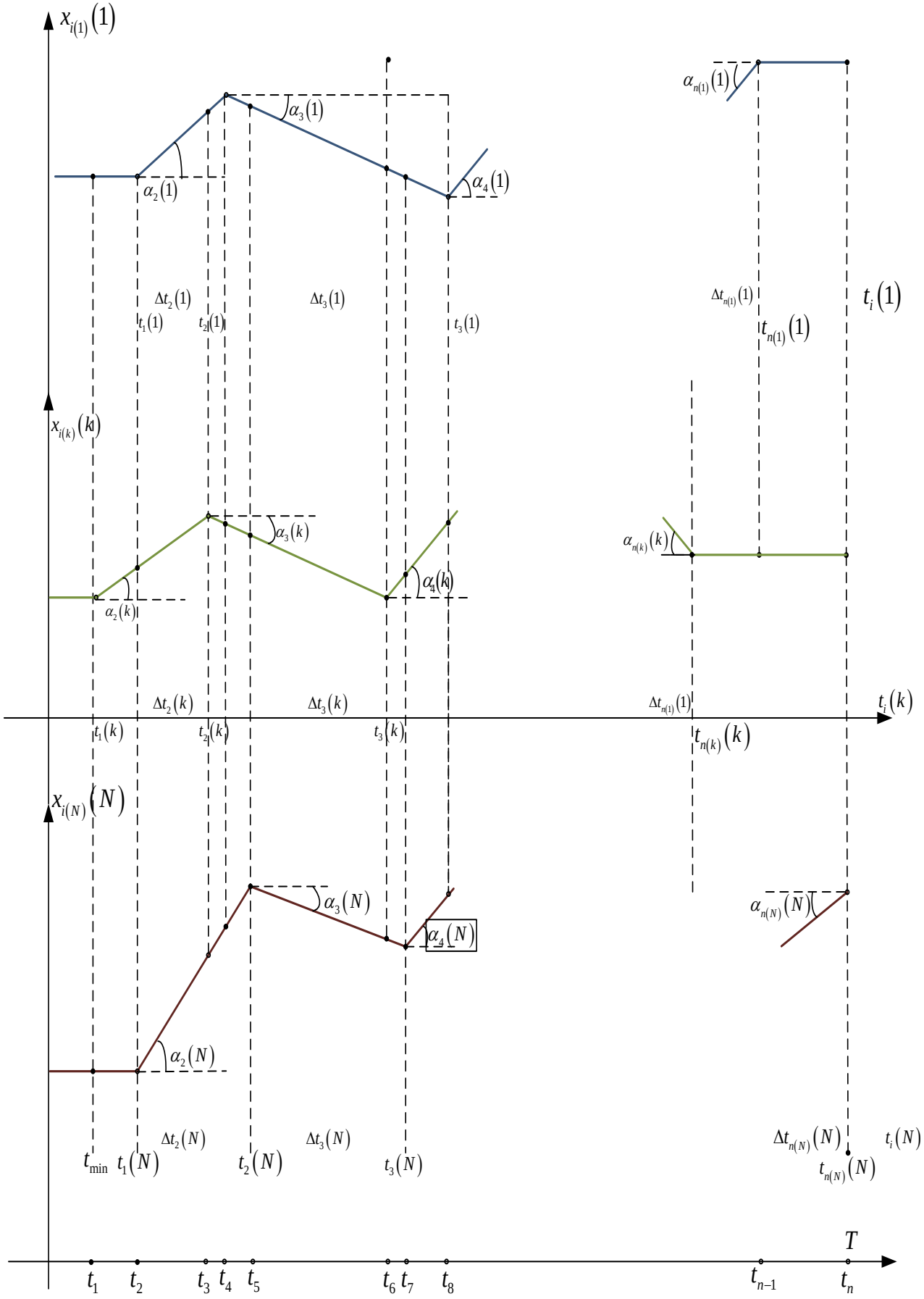

Рисунок 2 – Часові ряди з нерівномірними асинхронними тактами квантування

Використавши цільову функцію з налаштовними параметрами $\alpha, \gamma_{ji}, t,$ що забезпечує компроміс між стандартним FCM та методами кластеризації із зважуванням даних:

$$E^{KB}\left(u_j(k), \tilde{c}_j\right) = \sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2(k) \sum_{i=1}^{n}\left(\tilde{x}_i(k) - \tilde{c}_{ji}\right)^2 + \right.$$

$$\left. + (1-\alpha) u_j(k) \sum_{i=1}^{n} \gamma_{ji}^t \left(\tilde{x}_i(k) - \tilde{c}_{ji}\right)^2 \right) = \tag{15}$$

$$= \sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2(k) \left\| \tilde{x}(k) - \tilde{c}_j \right\|^2 + (1-\alpha) u_j(k) \left\| \tilde{x}(k) - \tilde{c}_j \right\|_{\Gamma_j^t}^2 \right),$$

де:

$$\Gamma_j = diag\left(\gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jn}\right), \mathrm{Tr}\Gamma_j = 1, \tag{16}$$

отримуємо метод кластеризації:

$$\begin{cases} u_j(k) = \dfrac{1 + \dfrac{(1-\alpha)}{2\alpha} \displaystyle\sum_{j=1}^{m} \dfrac{\left\| \tilde{x}(k) - \tilde{c}_j \right\|_{\Gamma_j^t}^2}{\left\| \tilde{x}(k) - \tilde{c}_j \right\|^2} - (1-\alpha) \left\| \tilde{x}(k) - \tilde{c}_j \right\|_{\Gamma_j^t}^2}{\dfrac{1}{2\alpha} \displaystyle\sum_{k=1}^{m} \left\| \tilde{x}(k) - \tilde{c}_j \right\|^{-2}} \Big/ 2\alpha \left\| \tilde{x}(k) - \tilde{c}_j \right\|^{-2}, \\ \gamma_{ji} = \left(\displaystyle\sum_{l=1}^{n} \left(\frac{\displaystyle\sum_{k=1}^{N} u_j(k) \left(\tilde{x}_i(k) - \tilde{c}_{ji}\right)^2}{\displaystyle\sum_{l=1}^{N} u_j(k) \left(\tilde{x}_l(k) - \tilde{c}_{jl}\right)^2} \right)^{\frac{1}{t-1}} \right)^{-1}, \\ \tilde{c}_{ji} = \dfrac{\displaystyle\sum_{k=1}^{N} \left(\alpha u_j^2(k) + (1-\alpha) u_j(k) \gamma_{ji}^t \right) \tilde{x}_i(k)}{\displaystyle\sum_{k=1}^{N} \left(\alpha u_j^2(k) + (1-\alpha) u_j(k) \gamma_{ji}^t \right)}. \end{cases} \tag{17}$$

Співвідношення (17) є узагальненням алгоритмів нечіткої кластеризації, заснованих на цільових функціях, які при $\alpha = 1$ перетворюються у стандартний FCM метод.

У **п'ятому розділі** проведено імітаційне моделювання методів навчання робастних адаптивних моделей часових рядів; імітаційне моделювання методів адаптивної можливісної нечіткої кластеризації часових рядів; послідовної онлайн нечіткої кластеризації багатовимірних рядів. Розв'язано практичну задачу на базі розроблених методів кластеризації у сучасних моніторингових системах.

Для подолання негативних факторів збуреності та аномальних викидів був використаний адаптивний робастний метод передобробки часових рядів в онлайн режимі надходження потоку даних. В роботі реалізовано ансамбль адаптивних моделей, який навчається в процесі обробки потоку даних. На рисунку 3 представлено архітектуру ансамблю адаптивних моделей ідентифікації.

В блоці селекції ансамблю адаптивних гібридних моделей в кожен момент дискретного часу проводиться вибір найкращого вихідного сигналу в рамках прийнятого критерію $MAPE(k) = \min_i [MAPE_i(k)]$. У таблиці 1 наведено результати порівняння значення на основі різних підходів. Запропонований підхід дозволяє обробляти сигнали в умовах істотного забруднення викидами.

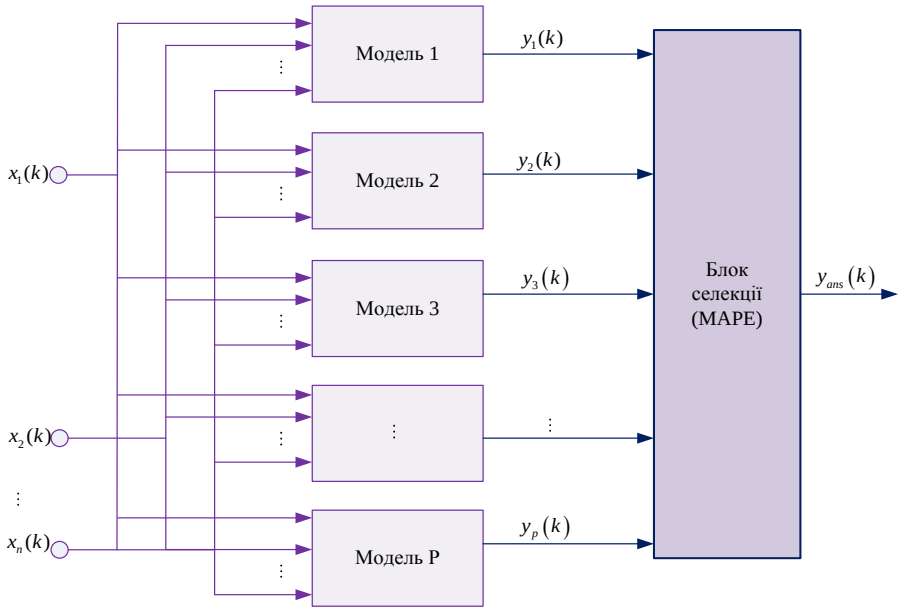

Рисунок 3 – Ансамбль адаптивних моделей ідентифікації

Таблиця 1 – Порівняння результатів ідентифікації зашумленних сигналів

Модель та метод навчання	RMSE
Налаштовна модель, навчана на основі алгоритму Гудвіна–Ремеджа–Кейнеса	0.3203
Налаштовна модель, навчана на основі модифікованої функції Гемана – МакКлюра	0.1932
Налаштовна модель, навчана на основі функції втрат Коші	0,0723
Налаштовна модель, навчана на основі рекурентного МНК	∞
Ансамбль гібридних адаптивних моделей ідентифікації	0.0735

З метою підтвердження ефективності адаптивної можливісної нечіткої онлайн кластеризації коротких часових рядів була вирішена задача кластеризації часових рядів погодинного споживання енергії. Результати запропонованого підходу дозволяють підвищити якість аналізу і прогнозування часових рядів. У якості критерію оцінювання результатів кластеризації була прийнята середня помилка класу.

Результати порівнянь наведені в таблиці 2. Вони представляють відсоток неправильно класифікованих об'єктів з набору даних тестування.

Таблиця 2 - Результати кластеризації часової серії

Методи кластеризації	M{СПК}
Алгоритм нечіткої ймовірнісної кластеризації	1.6 % (5)
Алгоритм адаптивної нечіткої ймовірнісної кластеризації	1.3 % (4)
Алгоритм можливісної кластеризації	11.1 % (33)
Адаптивний алгоритм можливісної кластеризації	6.3 % (19)

Розроблений метод обробки багатовимірних часових рядів з нерівномірним тактом квантування дозволив розв’язати задачу кластеризації даних у сучасних моніторингових системах. Для вирішення практичних задач кластеризації часових рядів, які послідовно надходять на обробку з нерівномірними тактами квантування, був розроблений модуль, у якому реалізовані методи адаптивної нечіткої кластеризації часових рядів, що базуються на апараті гібридних систем обчислювального інтелекту. За експериментальні дані були обрані медичні ряди відновної терапії після операції коронарного шунтування, які і є багатовимірними часовими рядами з нерівномірним тактом квантування. При порівнянні результатів з відомими методами та розрахунку значень для оцінки якості кластеризації був використаний індекс Девіса-Болдуїна. Результати показані у таблиці 3.

Таблиця 3 – Результати розрахунку значень для оцінки якості кластеризації

Індекси / кластерізаця	Індекс силуета	Індекс Девіса-Болдуїна
Алгоритм адаптивної нечіткої ймовірнісної кластеризації	0.2326	1.28
Алгоритм нечітких *c*-середніх	0.2354	1.23
Алгоритм *k*-середніх	0.3676	1.09
Алгоритм агломеративної кластеризації	0.2790	1.08
Mini Batch k-means	0.1904	1.50

Запропонований метод може бути використаний та орієнтований на послідовну адаптивну онлайн обробку часових рядів, що може забезпечити підвищення якості обробки забрудненої спотвореної інформації в умовах її багатовимірності та асинхронності, оскільки відомі аналоги орієнтовні на обробку даних в пакетному режимі.

У висновках сформульовано наукові та практичні результати, що одержано у дисертаційній роботі. У додатку наведено акти про впровадження результатів дослідження.

ВИСНОВКИ

У дисертаційній роботі розв’язана науково-практична задача нечіткої кластеризації нерівномірно квантованих часових рядів в інтелектуальному аналізі потоків даних. Створені методи дозволяють підвищити швидкодію обробки даних, що потрапляють на обробку в онлайн режимі. Внаслідок виконання роботи були

отримані нові наукові та практичні результати. Проведені дослідження дозволили зробити наступні висновки:

– досліджено моделі і методи інтелектуального аналізу даних та проведено загальний аналіз існуючих методів і підходів до кластеризації часових рядів; доведено, що існуючі методи не можуть обробляти часові ряди у онлайн режимі;

– вирішено задачу передобробки часових рядів, що спотворено аномальними викидами та розглянуто і запропоновано метод на основі модифікованої робастної функції, що має змогу обробляти часові ряди в онлайн режимі;

– розроблено онлайн метод кластеризації часових рядів за допомогою модифікації методів ймовірнісної та можливісної кластеризації для обробки коротких часових рядів в інтелектуальному аналізі даних;

– вперше розроблено послідовний метод онлайн кластеризації багатовимірних часових рядів за допомогою модифікованого підходу та введення метрики спеціального вигляду, що дозволило обробляти часові ряди високої розмірності, які послідовно надходять на обробку;

– вперше розроблено метод онлайн кластеризації асинхронних часових рядів, несхильних до впливу ефекту концентрації норм, що дозволяє вирішувати задачу кластеризації за умов перетину класів, за рахунок використання спеціальної цільової функції;

– на основі запропонованих методів проведено експерименти на основі тестових та реальних даних. Розроблено програмний модуль, що був застосований для моніторингу артеріального тиску захворювань серцево-судинної системи у період відновної терапії після операції коронарного шунтування;

– розроблені методи було програмно реалізовано та використано для ряду практичних впроваджень. Зокрема реалізовано модуль, що підтвердив свою ефективність у задачах моніторингу медичних даних в онлайн режимі. Результати досліджень впроваджені у ТОВ «Інфобуд», м. Харків (акт впровадження від 03.10.2018) та у ТОВ «Сайтосс», м. Харків (акт впровадження від 06.10.2018). Окремі положення, висновки та рекомендації дисертаційної роботи використано в освітньому процесі у Харківському національному університеті радіоелектроніки на кафедрі штучного інтелекту в курсі «Нейромережеві методи обчислювального інтелекту».

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Setlak, G., Bodyanskiy, Y., Pliss, I., Vynokurova, O., Peleshko, D., & Kobylin, I. (2017). Adaptive Fuzzy Clustering of Multivariate Short Time Series with Unevenly Distributed Observations Based on Matrix Neuro-Fuzzy Self-Organizing Network. In *Advances in Fuzzy Logic and Technology 2017* (pp. 308-315). Springer, Cham. (Входить до міжнародної наукометричної бази SCOPUS).

2. Bodyanskiy, Y., Vynokurova, O., Kobylin, I., & Kobylin, O. (2016). Adaptive Fuzzy Clustering of Short Time Series with Unevenly Distributed Observations in Data Stream Mining Tasks. *Information Technology and Management Science*, *19*(1), 23-28. (Входить до наукометричної бази SCOPUS).

3. Бодянский, Е. В., Винокурова, Е. А., Кобылин, И. О., & Мулеса, П. П. (2016). Робастная адаптивная идентификация нестационарных временных рядов с помощью ансамбля обучаемых гибридных адаптивных моделей. *Управляющие системы и машины*, (5), 76-83.

4. Бодянський, Є., Винокурова, О., Кобилін, І., & Мулеса, П. (2017). Адаптивна матрична нейро-фаззі самоорганізовна мережа для кластеризації багатовимірних потоків даних. *Вісник Національного університету «Львівська політехніка». Серія: Комп'ютерні науки та інформаційні технології*, (864), 314-319.

5. Бодянский, Е.В., Винокурова, Е.А., Кобылин, И.О., Кобылин, О.А., & Пелешко, Д.Д. (2017) Нечёткая кластеризация временных рядов с неравномерными и асинхронными тактами квантования. *Системи обробки інформації, 5(151)*, 47-54.

6. Bodyanskiy, Y., Vynokurova, O., Szymański, Z., Kobylin, I., & Kobylin, O. (2016, August). Adaptive Robust Models for Identification of Nonstationary Systems in Data Stream Mining Tasks. In *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)* (pp. 263-268). IEEE. (Входить до наукометричної бази SCOPUS).

7. Bodyanskiy, Y., Kobylin, I., Rashkevych, Y., Vynokurova, O., & Peleshko, D. (2018, February). Hybrid Fuzzy-Clustering Algorithm of Unevenly and Asynchronously Spaced Time Series in Computer Engineering. In *2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET)* (pp. 930-935). IEEE. (Входить до міжнародної наукометричної бази SCOPUS).

8. Бодянский, Е. В., Дейнеко, А. А., Кобылин, И. О., & Плисс, И. П. (2016). Адаптивная нечеткая кластеризация коротких временных рядов в интеллектуальном анализе потоков данных. *Intellectual Systems For Decision Making and Problems of Computational Intelligence*, 255-257.

9. Бодянський, Є. В., Винокурова, О. А., Ізонін, І. В., Кобилін, І. О., & Мулеса, П. П. (2017) Кластеризація багатовимірних часових рядів на основі адаптивної матричної нейро-фаззі самоорганізовної мережі. *Intellectual Systems For Decision Making and Problems of Computational Intelligence*, 247-248.

10. Бодянський, Є. В., Винокурова, О. А., Кобилін, І. О., & Мулеса, П.П. (2016). Адаптивна нечітка кластеризація багатовимірних часових рядів з нерівномірним тактом квантування. *Праці VIII-Й Міжнародної школи семінару-«Теорія Прийняття Рішень»* 56-57.

11. Кобылин, И.О., (2015) Об одном методе кластеризации коротких временных рядов. *"Радиоэлектроника и молодежь в XXI веке"* 30-31.

12. Кобылин, И.О., (2016) Адаптивная кластеризация коротких временных рядов с неравномерным тактом квантования. *"Радиоэлектроника и молодежь в XXI веке"* 21-22.

АНОТАЦІЯ

Кобилін І.О. Нечітка кластеризація часових рядів в інтелектуальному аналізі потоків даних. – На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.23 - системи та засоби штучного інтелекту. – Харківський національний університет радіоелектроніки, Міністерство освіти і науки України, Харків, 2019.

На сьогоднішній час тенденція обробки великих обсягів інформації та їх аналіз за допомогою кластеризації дає змогу зрозуміти різноманіття процесів для її подальшого використання у сферах життєдіяльності, що супроводжують людину.

Значну частину інформації, пов'язану з обробкою великих обсягів даних, містять часові ряди. Однак, однією з типових проблем обробки часових рядів є їх нерівномірне квантування та багатовимірність.

Предметом дослідження є методи інтелектуального аналізу для нечіткої онлайн кластеризації багатовимірних часових рядів з нерівномірними та асинхронними тактами квантування, що призначені для аналізу потоків даних.

Розглянуто питання розвитку та використання методів нечіткої кластеризації в умовах недостатньої кількості спостережень та забрудненості оброблюваних даних. Наведено огляд технологій для вирішення задач класифікації, кластеризації та фільтрації.

Запропоновано метод нечіткої кластеризації, який ефективно працює за умов перетину класів та несхильний до ефекту концентрації норм та працює в онлайн режимі з асинхронними нерівномірно квантованими часовими рядами за рахунок використання спеціальної цільової функції. Запроваджений метод може бути корисний при вирішенні завдань, що виникають в рамках інтелектуального аналізу потоків даних, коли вихідні дані мають високу розмірність.

Запропоновано послідовний онлайн метод кластеризації багатовимірних часових рядів, що базується на апараті гібридних систем обчислювального інтелекту, який дозволяє вирішувати задачу кластеризації даних, які послідовно надходять на обробку, з нерівномірними тактами квантування.

Розроблено метод адаптивної ймовірнісної та можливісної кластеризації, що базується на метриці спеціального вигляду, в основі якої лежить аналіз тангенсів кутів нахилу часового ряду, що дозволило спростити чисельну реалізацію методу та розв'язувати задачу кластеризації нерівномірно квантованих часових рядів і формалізувати розв'язання задачі нечіткої кластеризації коротких часових рядів. Відмінною особливістю методики є оцінка якості кожного розбиття і вибір найкращого з них.

Запропоновано метод робастної адаптивної ідентифікації нестаціонарних часових рядів, що в онлайн режимі надходять на обробку, який характеризується простотою обчислювальної реалізації та стійкістю до аномальних викидів. Розглянутий метод простий у чисельній реалізації, будучи по суті градієнтним методом оптимізації цільових функцій спеціального виду.

Проведено низку імітаційних експериментів на основі тестових та реальних даних, результати яких підтверджують доцільність застосування запропонованого підходу для вирішення задач інтелектуального аналізу потоків даних. Розв'язано практичну задачу на базі розроблених методів кластеризації для вирішення медичної проблеми захворювання серцево-судинної системи.

Ключові слова: часові ряди, нечітка кластеризація часових рядів, асинхронність квантування, онлайн метод нечіткої кластеризації, адаптивні методи навчання, інтелектуальний аналіз даних, робастні цільові функції, нестаціонарні нелінійні часові ряди, ймовірнісна кластеризація, можливісна кластеризація.

АННОТАЦИЯ

Кобылин И.О. Нечеткая кластеризация временных рядов в интеллектуальном анализе потоков данных. – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.23 – системы и средства искусственного интеллекта. – Харьковский национальный университет радиоэлектроники, Министерство образования и науки Украины, Харьков, 2019.

Предложен метод нечеткой кластеризации, который эффективно работает в условиях пересечения классов, не подвержен эффекту концентрации норм в онлайн режиме с асинхронными неравномерно квантованными временными рядами за счет использования специальной целевой функции.

Предложен последовательный онлайн метод кластеризации многомерных временных рядов, основанный на аппарате гибридных систем вычислительного интеллекта, позволивший решать задачу кластеризации данных, которые последовательно поступают на обработку с неравномерными тактами квантования.

Разработан метод адаптивной кластеризации на основе вероятностной и возможностной кластеризации, сформированный на метрике специального вида, в основе которой лежит анализ тангенсов углов наклона временного ряда, что позволило упростить численную реализацию метода и решать задачу кластеризации неравномерно квантованных временных рядов.

Предложена модель робастной адаптивной идентификации нестационарных временных рядов в онлайн режиме поступления потока данных, который характеризуется простотой вычислительной реализации и устойчивостью к аномальным выбросам.

Ключевые слова: временные ряды, нечеткая кластеризация временных рядов, асинхронность квантования, онлайн методы нечеткой кластеризации, адаптивные методы обучения, интеллектуальный анализ данных, робастные целевые функции, нестационарные нелинейные временные ряды, вероятностная кластеризация, возможностная кластеризация.

ABSTRACT

Kobylin I.O. Fuzzy clustering of time series in data stream mining. – Manuscript copyright.

Dissertation for obtaining the scientific degree of the candidate of technical sciences on the specialty 05.13.23 – Systems and tools of artificial intelligence. – Kharkiv National University of Radio Electronics, Ministry of Education and Science of Ukraine, Kharkiv, 2019.

The offered fuzzy clustering method works effectively under conditions of classes overlapping and is free from the norms concentration effect under online mode with asynchronous non-stationary quantized time series through a special objective function.

The offered sequential online clustering method for multidimensional time series, based on the apparatus of hybrid systems of computational intelligence, enables solves the problem of clustering data that are sequentially send to processing with non-stationary quantization cycles.

The developed method of adaptive modification based on probabilistic and possibilistic clustering based on a special type metric stemming from slope ratio analysis of the time series, that simplifies the numerical implementation of the methods and solves the clustering problem for non-stationary quantized time series.

The offered model of robust adaptive identification of non-stationary time series, under online mode, of data flow, characterized by simple computational implementation and resistance to anomalous emissions.

Key words: time series, fuzzy clustering of time series, asynchronous quantization, online fuzzy clustering procedure, adaptive learning procedures, data stream mining, robust objective functions, nonstationary time series, probabilistic clustering, possibilistic clustering.